Drug combination prediction for cancer treatment using disease-specific drug response profiles and single-cell transcriptional signatures.

Daniel Osorio[1,✉], **Xavier Tekpli**[2], **Vessela N. Kristensen**[2,3], **and Marieke L. Kuijjer**[1,4,5,✉]

[1]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway.
[2]Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway.
[3]Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway
[4]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
[5]Leiden Center for Computational Oncology, Leiden University Medical Center (LUMC), Leiden, The Netherlands

Developing novel cancer treatments is a challenging task that can benefit from computational techniques matching transcriptional signatures to large-scale drug response data. Here, we present '*retriever*,' a tool that extracts robust disease-specific transcriptional drug response profiles based on cellular response profiles to hundreds of compounds from the LINCS-L1000 project. We used *retriever* to extract transcriptional drug response signatures of triple-negative breast cancer (TNBC) cell lines and combined these with a single-cell RNA-seq breast cancer atlas to predict drug combinations that antagonize TNBC-specific disease signatures. After systematically testing 152 drug response profiles and 11,476 drug combinations, we identified the combination of kinase inhibitors QL-XII-47 and GSK-690693 as the topmost promising candidate for TNBC treatment. Our new computational approach allows the identification of drugs and drug combinations targeting specific tumor cell types and subpopulations in individual patients. It is, therefore, highly suitable for the development of new personalized cancer treatment strategies.

drug-combination | cancer | single-cell RNA-seq | LINCS-L1000 | drug re-purposing

Correspondence: ***daniecos@uio.no*** & ***marieke.kuijjer@ncmm.uio.no***

Introduction

Developing new drugs and pharmacological regimens to treat complex diseases such as cancer is an expensive and challenging task (1). Several computational techniques that use structural interactions (2), transcriptional signatures (3–5), biological networks perturbations (6, 7), and data mining (8) are available to aid in the discovery of new treatment regimens which could improve patient management (9). Methods based on transcriptional signatures use the observed changes in gene expression profiles between samples from patients affected with the disease under study compared to samples from control subjects, and match these to the response profiles of cell lines that act as surrogates for the disease to different compounds. These data can be combined to identify compounds that may antagonize the disease's transcriptional changes, returning it to a more healthy-like state (5). Finding specific transcriptional response profiles to drug candidates, as well as robust disease-associated transcriptional alterations, is therefore crucial for such approaches to produce reliable predictions.

To obtain precise disease-specific transcriptional signatures that can be targeted by pharmacological compounds, one needs to account for transcriptional variability between patients by including samples from multiple individuals (10). This allows for the identification of a target set of genes that is constitutively expressed with the same magnitude and pattern across patients/cells. Targeting such a consistently expressed set of genes is thought to increase the effectiveness and safety of the treatment (11). Transcriptional signatures are measurable at a variety of resolutions, ranging from low resolution, as in the complex mixture of cells present within tissues, to high resolution, such as single-cells (12). The Cancer Genome Atlas (TCGA) project provides RNA-sequencing (RNA-seq) data for tumors and adjacent tissues (13). While such tissue-level data have been useful in the identification of cancer subtypes and prognostic signatures, the identification of accurate transcriptional signatures between disease and healthy states has been hindered by the wide heterogeneity of cell types found in tumor tissues (14).

This issue can be overcome by using single-cell RNA-seq data (15). A drawback of single-cell RNA-seq experiments is that, due to the higher cost of profiling of single cells, often fewer biological replicates are available. However, by combining multiple experiments, sample sizes of cells with a desired molecular phenotype can be increased, improving characterization of the transcriptional changes observed in disease (16). In addition, the construction of pseudo-bulk profiles (the sum of the expression values of a gene across all cells obtained from the same individual) from single-cell data makes it possible to account for gene expression variability among patients (17, 18).

The LINCS-L1000 project published transcriptional profiles of several cell lines, treated with hundreds of compounds at various concentrations and time points of drug exposure (4). These profiles have previously been used to identify potential drugs that can be repurposed to treat a variety of diseases, including cancer (3). Additionally, the project provides an interactive portal in which users can interrogate whether an up- or down-regulated gene set of interest overlaps with transcriptional drug response signatures. The portal then returns a ranked list of compounds that are likely to have an inverse effect on disease-associated gene expression levels (19). However, these predictions are not based on robust tissue- or disease-specific transcriptional profiles and may therefore over- or underestimate the potential effect of drugs on specific diseases. Additionally, the LINCS-L1000 web portal returns an exhaustive list of independent experiments with matching inverse patterns, rendering it difficult to identify compounds that induce stable transcriptional responses in cell lines derived from the same disease across multiple time points and drug concentrations. Thus, being able to extract robust disease-specific transcriptional drug response signatures that are consistent at different time points, drug concentration, or cell line from the profiles provided by the LINCS-L1000 project would significantly improve drug prioritization and accelerate the identification of new pharmacological options for personalized treatment of cancer (20).

Once both the disease profile and the disease-specific response to multiple compounds are available, rank-based correlation analysis can be used to quantitatively identify compounds that can revert the transcriptional changes that distinguish diseased samples from healthy ones (5, 9). Nonetheless, monotherapy in cancer is highly susceptible to the development of resistance following an initial response to treatment (21). Combination therapy, or the simultaneous administration of multiple drugs to treat a disease, has evolved into the standard pharmacological regimen for treating complex diseases such as cancer. Combination therapy prevent tumor evolution and help inhibit the development of drug resistance in cancer, thereby improving patient survival (22).

The *in silico* prediction of responses to drug combinations is an active topic of research in computational biology (23–25). Recently, Pickering (2021) found that the transcriptional response signatures to $856,086$ unique two-drug combinations could be predicted based on the $1,309$ drug response profiles to compounds tested by the Connectivity Map Build 2 project (parent of the LINCS-L1000 project) (26). After analyzing 148 independent studies involving 257 treatment combinations from the Gene Expression Omnibus (GEO) database, it was discovered that averaging the expression profiles of individual treatments provides 78.96 percent accuracy in predicting the direction of differential expression for the combined treatment (27). However, different from the Connectivity Map project—where the generation of combinatorial response profiles is possible thanks to the availability of single response profiles for each individual compound (28)—the LINCS-L1000 project does not provide robust single drug response profiles, making the generation of meaningful combinatorial profiles not yet possible.

Here, we present *retriever*, a tool that uses correlation analysis and hierarchical collapsing to extract robust disease-specific transcriptional drug response signature profiles that are consistent across time, concentration, and cell line, from data provided by the LINCS-L1000 project. We integrated these transcriptional drug response profiles with a single-cell RNA-seq signature representing the transcriptional changes observed in triple-negative breast cancer cells compared to healthy breast epithelial cells. We used these two signatures—the transcriptional changes induced by the disease and the cellular responses to drugs—to prioritize drugs and to predict novel drug combinations suitable for the treatment of triple-negative breast cancer. The transcriptional changes associated with TNBC were computed from a single-cell breast cancer RNA-seq atlas built by us after compiling single-cell RNA-seq data obtained from 36 publicly available healthy breast and breast cancer samples. After systematically testing drug response profiles and drug combinations, we identified the combination of kinase inhibitors QL-XII-47 and GSK-690693 as the topmost promising treatment for TNBC. In addition to recommending drug combinations, the profiles returned by *retriever* allow for the characterization of possible mechanisms of action of the identified compound(s) to reverse the disease's transcriptional profile towards a healthy-like state.

Methods

Single-cell RNA-seq datasets. We collected publicly available single-cell RNA-seq count matrices for healthy breast tissues and breast cancer samples from multiple sources (see Data Availability). Each sample's gene identifiers (IDs) were translated into current gene symbols using the dictionary of gene ID synonyms provided by ENSEMBL (29). Datasets were loaded into R and integrated into a single 'Seurat' object (30). Data were then subjected to quality control keeping only cells with a library size of at least $1,000$ counts and within the 95 percent confidence interval of the prediction of the mitochondrial content ratio and detected genes in proportion to the cell's library size. We also removed all cells that had mitochondrial proportions greater than 10% (31). We then normalized, scaled, and reduced the dimensionality of the data through Principal Component Analysis (PCA) using the default functions and parameters included in Seurat. Data integration was performed using Harmony (32). UMAP projections of the data were extracted using the top 20 dimensions returned by Harmony (33). Cell clustering was performed using the functions included in Seurat for this purpose, with a resolution of 0.01, using UMAP embeddings as the source for the nearest neighbor network construction.

Cell type assignation. Assignation of cell types was performed based on the expression of the marker genes reported by Wu et al. (34, 35). Epithelial cell subtypes were assigned using the Nebulosa package by querying cells expressing *ESR1*, *PGR*, and *ERBB2* receptors (36).

Single-cell RNA-seq differential expression analysis. Transcriptional signatures of triple-negative breast cancer epithelial cells compared to healthy epithelial cells were quantified by differential expression using MAST (37). We compared (Supplementary Fig. S1 and Fig. S2) all epithelial cells from cancer patients ($5,730$ cells, $31,73\%$) against those from healthy donors ($12,323$ cells, 68.26%). In addition, we compared the cluster of patient-derived epithelial cells depleted of hormone and growth factor receptor expression (enrichment for $ESR1^{-}$, PGR^{-}, and $ERBB2^{-}$ cells—$2,998$ cells, 16.6%) against all healthy epithelial cells ($12,323$ cells, 68.3%), against the cluster of epithelial cells depleted of hormone and growth factor expression (6117 cells, 33.9%) derived from healthy individuals, and against the cluster of epithelial cells enriched for triple-positive cells ($ESR1^{+}$, PGR^{+}, and $ERBB2^{+}$) from healthy individuals ($6,206$ cells, 34.4%). We selected the comparison of the cluster enriched with triple-negative cancer cells with the cluster enriched for triple-positive healthy epithelial cells as the disease-specific transcriptional profile, as this profile showed the highest Spearman Correlation Coefficient ($\hat{\rho}$) with tissue-level fold-changes computed on TCGA breast cancer data.

Pseudo-bulk analyses. Pseudo-bulk profiles were computed by summing up all expression values for the same gene across cells of the same individual. We applied this procedure to the triple-negative epithelial cells derived from patients against the triple-positive epithelial cells from healthy individuals. The computed pseudo-bulk profiles from diseased and healthy individuals were then compared using DESeq2 (38).

TCGA breast cancer dataset analysis. Upper-quantile FPKM-transformed RNA-seq data at the tissue level as well as the associated metadata were downloaded from the TCGA breast cancer (TCGA-BRCA) project using the 'TCGAbiolinks' Bioconductor package (39). Samples labeled as 'healthy' and 'basal' were collected and differential expression between these groups was evaluated using DESeq2 (38).

Comparisons between data-types. The $\log_2$ fold-changes of expression between cancer and control samples at the single-cell level, pseudo-bulk level and tissue level were used to validate the computed signatures describing the transcriptional changes in triple-negative breast cancer compared to healthy epithelial cells using the Spearman correlation coefficient (40).

Application of *retriever* to triple-negative breast cancer. The transcriptional response profiles of triple-negative breast cancer cell lines that were treated with different concentrations and treatment lengths of different compounds available in the LINCS-L1000 project were collected using the 'ccdata' Bioconductor package (28). The LINCS-L1000 project provides these profiles for three different TNBC cell lines (BT20, HS578T, and MDAMB231). We used these data to compute the disease-specific transcriptional response profiles.

Generation of drug combination transcriptional response profiles. All possible two-drug combinations were generated by averaging the computed disease-specific transcriptional response profiles.

Drug repurposing analysis. Using the set of overlapping genes between the LINCS-L1000 project and differential expression analysis from single-cell RNA-seq data, we performed Spearman correlation analyses with the independent disease-specific drug response profiles, as well as with the drug combination profiles. Ninety-five percent confidence intervals, p-values, and false-discovery rates (FDR) (41) were also computed. Compounds were ranked based on the computed correlation coefficients, from negative to positive.

Mechanisms of action prediction. To identify possible mechanisms of action for drugs or drug combinations, we took the disease-specific transcriptional response profiles and used these as input for GSEA, with the MSigDB Hallmarks gene set as reference to compute the enrichment of pathways as well as the directionality of the effect (42). Pathways with FDR < 0.05 were considered significant.

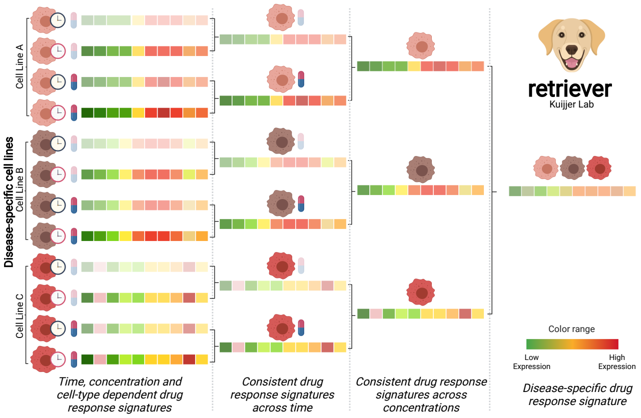
Cell Line A
Cell Line B
Cell Line C
Disease-specific cell lines
retriever
Kuijjer Lab
Color range
Low Expression
High Expression
Time, concentration and cell-type dependent drug response signatures
Consistent drug response signatures across time
Consistent drug response signatures across concentrations
Disease-specific drug response signature

Fig. 1. Overview of the *retriever* algorithm. *retriever* generates disease-specific transcriptional drug response signatures by merging transcriptional signatures over time, concentration, and cell-type. These signatures can then be matched to single-cell or bulk expression profiles to predict drugs and drug combinations most likely to be effective in treating a disease.

Results

The *retriever* algorithm. The *retriever* algorithm extracts disease-specific drug response profiles using three steps (Figure 1). The first step summarizes the cellular responses at different time points after the application of the drug, the second step summarizes the responses at different concentrations, and the third step summarizes the responses across different cell lines. This provides robust, disease-specific transcriptional response profiles based on the responses observed in all cell lines used as surrogates of a specific disease.

In the first step, *retriever* takes the response profiles of a given cell line to the same compound under the same concentration, but at different time points, and averages these. Then, the descriptive power of the extracted profile to represent the transcriptional response to a drug at a given concentration in the cell line, consistent across time, is evaluated using Spearman's correlation coefficient.

The averaged profile is returned if the correlation with the computed profile is larger than a user-defined threshold (default $\rho = 0.6$). This threshold can take values between 0 and 1, representing the percentage of agreement between the cellular responses to the same compound. If the threshold is too close to one, very few averaged profiles will be returned due to strong filtering, and if it is defined to close to zero, very noisy profiles (similar to simply averaging all the available profiles) will be returned. The original drug response profiles that do not reach this threshold are removed. The averaged profile is then recomputed using all response profiles that reached the threshold. This procedure ensures the removal of aberrant or insufficient cellular responses. Only averaged signatures of at least two profiles are used in the second step.

In the second step, *retriever* takes the stable time-consistent signature profiles of a compound at a particular concentration in the same cell line. To summarize the response at different concentrations, it applies the same procedure described in the first step over the averaged profiles. The profiles returned by the second step are stable transcriptional response profiles that are consistent across time and drug concentration in a specific cell line.

The last step extracts disease-specific drug response profiles by again applying the procedure described in step 1 to the stable response profiles to the compound in all disease-specific cell lines available in the LINCS-L1000 project. The profiles returned by the third step are robust disease-specific response signatures representing transcriptional changes to a specific compound.

Single-cell RNA-seq atlas of breast samples. To showcase how *retriever* can be used to prioritize drugs, we applied it to single-cell data from breast cancer and healthy breast tissues. For this, we compiled 36 publicly available single-cell RNA-seq count matrices from breast samples (26 diseased and 10 healthy). In total, we combined $109,097$ cells into a single Seurat object, maintaining the sample of origin metadata. Following quality control (see Methods section), $77,384$ cells were retained for further analysis (Figure 2A), of which $30,790$ were derived from healthy (left panel in Figure 2B) and $46,594$ from cancer samples (middle panel in Figure 2B). Cell types were assigned to the nine identified clusters in the low dimensional representation (Figures 2C, S3) using markers reported by Wu et al. (34, 35).

Transcriptional signatures associated with triple-negative breast cancer at the single-cell level. To identify the transcriptional signature that best represents the changes associated with triple-negative breast cancer, we first identified the single cells with a triple-negative phenotype within the epithelial cluster of cells. To do so, we queried cells expressing ESR1, PGR, and ERBB2 receptors, which allowed us to assign a cluster of cells enriched for the expression of the three receptors (Right panel in Figure 2B); cells belonging to the cluster enriched for triple-positive cells will be further referred as “triple-positive-like" cells in the remainder of this work. While cells belonging to the other cluster will be referred to as “triple-negative-like" cells.

We compared the triple-negative-like breast cancer cells with different healthy epithelial cell subpopulations, including all epithelial cells, cells belonging to the triple-negative cluster, and cells belonging to the triple-positive cluster of epithelial cells. We then compared fold changes to those derived from the population level-data from TCGA (see Methods section). The most representative profile was the one comparing triple-negative-like epithelial cancer cells with healthy triple-positive-like epithelial cells. Note that, while approximately half of all healthy epithelial cells are assigned to the triple-positive cluster, these receptors are expressed in only 20% of these cells, and at lower levels compared to breast cancer cells (Supplementary Fig. S1).

As a result, we used MAST to perform differential expression analysis at the single cell level between triple-negative-like epithelial cancer cells ($n = 2,998$) and healthy triple-positive-like epithelial cells ($n = 6,206$), identifying 205 differentially expressed genes (106 upregulated and 99 downregulated) with absolute $\log_2$ fold-changes larger than 1 and false discovery rate lower than 0.05 (Figure 3A).

After testing using both the hypergeometric test through Enrichr (43) and gene set enrichment analysis (GSEA) using the Hallmark MSigDB signatures as reference gene sets (42), we found that the differentially expressed genes were associated with the activation of *Oxidative Phosphorylation Pathway*, *Interferon Alpha Response*, as well as *Myc Targets*, and the downregulation of *TNFα Signaling via NF-κB*, *Estrogen Response*, *UV Response Up*, *Apoptosis*, *Hypoxia*, *Unfolded Protein Response*, *IL-6/JAK/STAT3 Signaling*, *Inflammatory Response*, and the *Androgen Response* pathway (Figure 3C, Supplementary Table S1). Many of these pathways are known to be highly associated with TNBC molecular phenotype. For example, activation of both the oxidative phosphorylation pathway and of Myc targets is associated with a worse outcome of the disease (44, 45), and the crosstalk between the interferon alpha response pathway and NF-κB is associated with drug resistance and tumor progression in this malignancy (46). The downregulation of the estrogen response pathway as well as apoptosis and hypoxia are markers of TNBC (47). In addition, downregulation of androgen and inflammatory responses is associated with worse prognosis and higher chemotherapy responsiveness respectively (48, 49).

We found $6,149$ genes expressed in both the single-cell RNA-seq datasets and the bulk RNA-seq from the TCGA-BRCA project. Spearman correlation coefficients between the expression changes computed at the single-cell level (Left panel in Figure 3A), pseudo-bulk level (Middle panel in Figure 3A) and tissue level (Right panel in Figure 3A), revealed a monotonic positive association between these different levels (Figure 3B), supporting the descriptive power of the computed differential single-cell expression signature to describe transcriptional changes associated with TNBC, and ensuring that the selected cell type as well as the pseudo-bulk samples are high-resolution descriptors of the diseased tissues.

Applying *retriever* to extract TNBC-specific transcriptional drug-response signatures. We collected $4,899$ response profiles measured in TNBC cell lines from the LINCS-L1000 project using the ‘ccdata’ package available in Bioconductor (28). These profiles correspond to the expression change of 1000 genes in response to 205 compounds on average at four different concentrations (0.08μM, 0.4μM, 2μM, and 10μM), and at two time points (6 and 24 hours), in three different TNBC cell lines (BT20, HS578T, and MDAMB231).
An illustration of the step-by-step process of constructing a generalized response profile to a compound across three TNBC cell lines is presented in Figure 4 using the QL-XII-47 compound as a case example.

1. *Computing time-consistent generalized response profiles:* In the first step (Figure 4A), we computed time-consistent generalized response profiles. To achieve this, we took the response profile of the MDAMB231 cell line to QL-XII-47 under the same concentration at two different time points (6 and 24 hours) and averaged them. When we compared

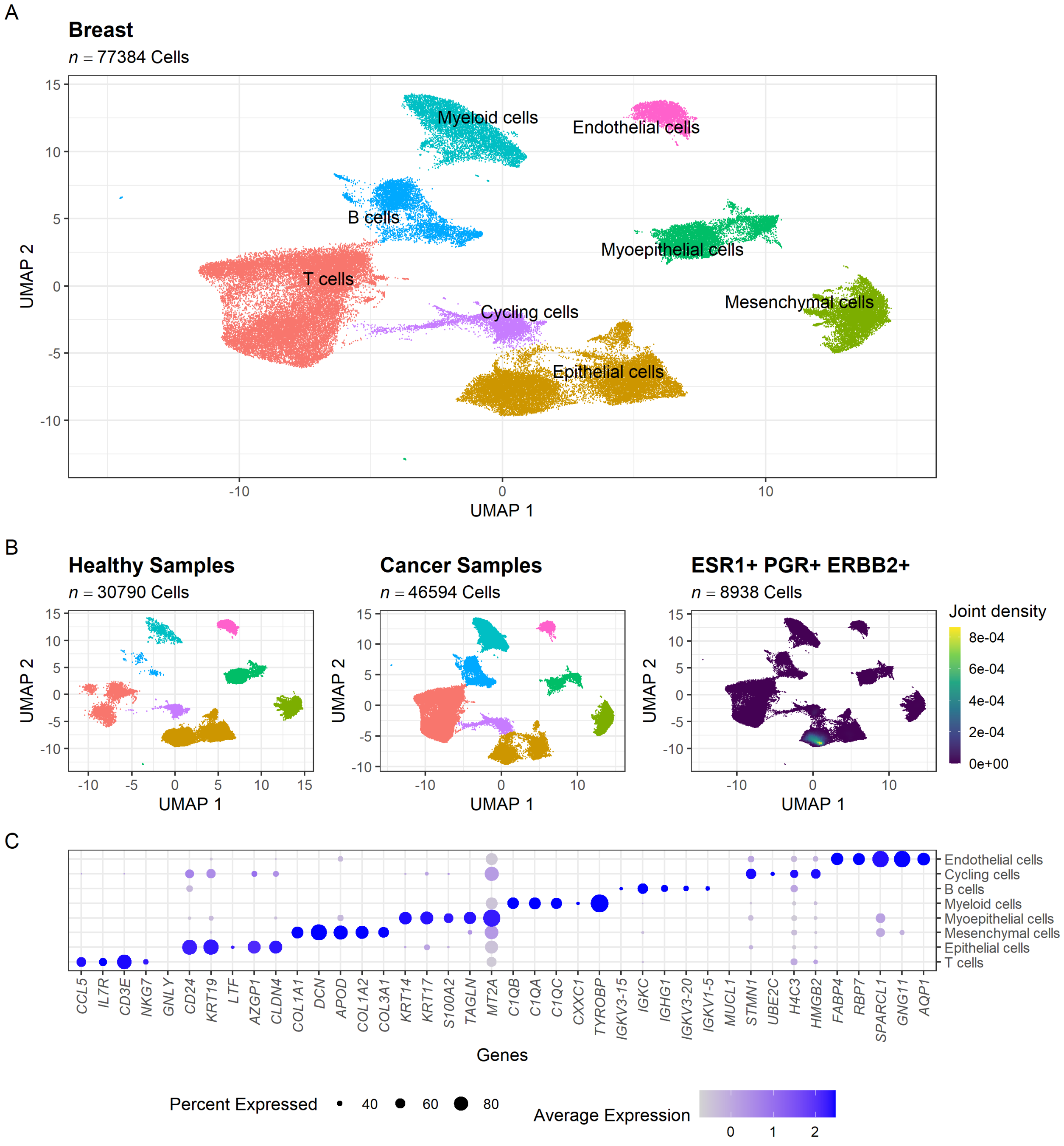

Fig. 2. Single-cell atlas of breast samples. **(A)** UMAP projection of the integrated 77,384 cells from 36 breast samples. **(B)** UMAP projection of the healthy (left) and cancer (center) cells, and visualization of triple-positive epithelial cells in the cancer samples. **(C)** Dotplot representing the normalized expression level and percentage of cells expressing the top five differentially expressed genes for each cell type.

the transcriptomes of the cell line exposed at two different time points, we found little or no correlation (large boxes labeled in gray) for each concentration (0.08μ M, 0.4μ M, 2μ M, and 10μ M). However, we found that their averaged profile displays high predictive power (Spearman Correlation Coefficient ($\hat{\rho}$) > 0.65 in all cases) of the cellular response to the compound, independent of the time of treatment. When present, we removed the profiles that did not correlate with the averaged profile above 0.6 (A total of 19.72% of the profiles). This procedure allows to identify concentrations in which the compounds induces a different, aberrant or insufficient cellular responses to the drug and also allows us to compute a single robust response profile to the compound.

2. *Computing time- and concentration-consistent response profiles:* In the second step (Figure 4B), we extract time and concentration-consistent responses. For this purpose, we averaged the time-consistent profiles computed in the first step. As before, we removed the profiles that did

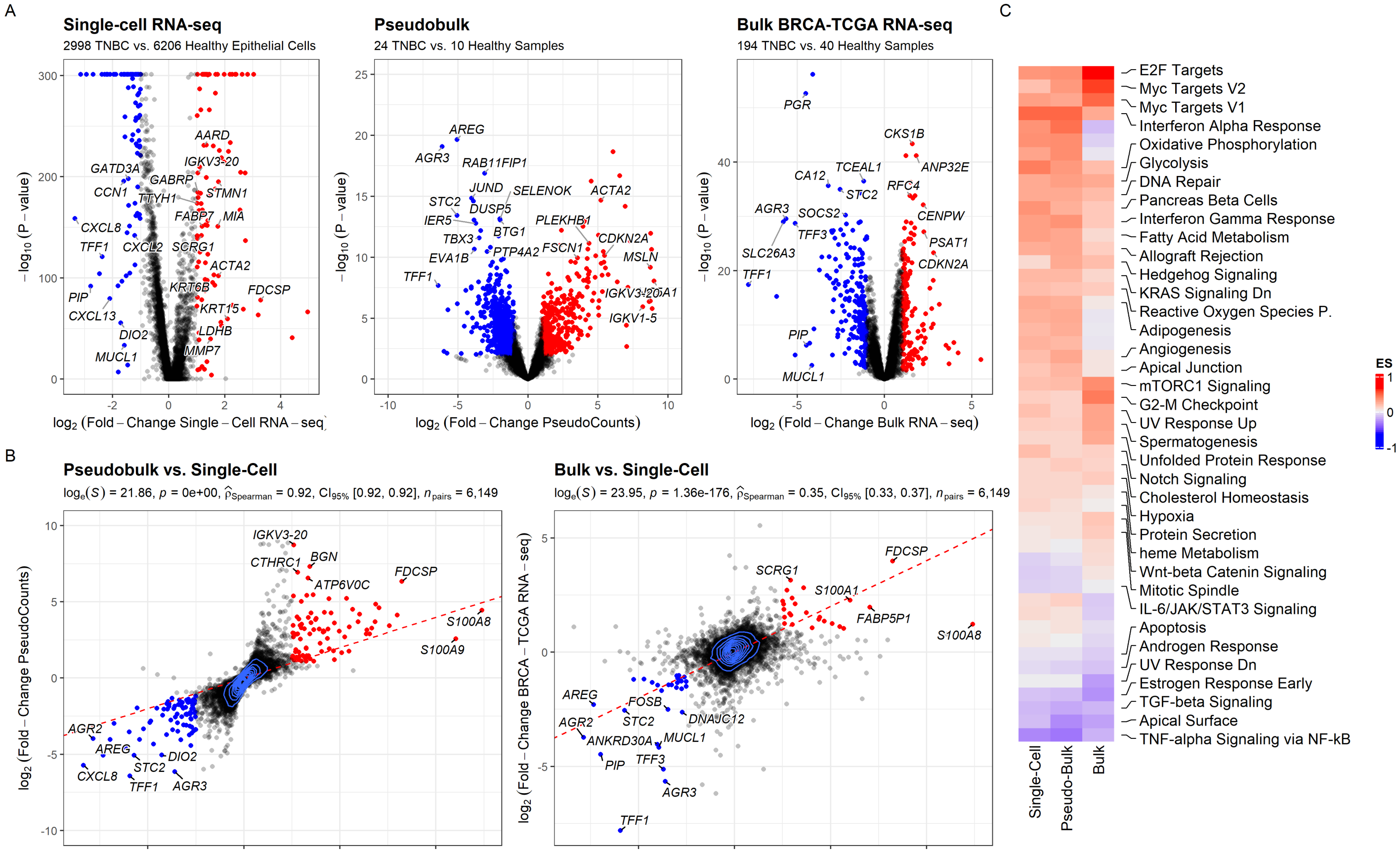

Fig. 3. Transcriptional changes identified in TNBC cells. **(A)** Volcano plots report the difference between TNBC and healthy samples, on the left based on single-cell RNA-seq count data computed using MAST, in the middle using the pseudo bulk measures in each single-cell RNA-seq sample, and on the right using the bulk RNA-seq data from the BRCA-TCGA project. Each dot represents a gene. Dots are color coded, in red if the $\log_2$ fold-change is larger than 1 and in blue if the $\log_2$ fold-change is smaller than -1. **(B)** Comparisons of the transcriptional changes associated with TNBC at the single-cell, pseudo bulk and tissue level. Each dot represents a gene. Dots are color coded as in Figure 2A. **(C)** Single-sample GSEA Enrichment Score (ES) for the transcriptional changes between TNBC and healthy cells at different levels of resolution. Labeled pathways are those that show same trend (positive or negative ES) in the three data types.

not correlate with the averaged profile above 0.6, such as the profile present in the first panel of Figure 4B, which showed a different cellular response at the concentration (A total of 16.91% of the profiles computed in Step 1). We then recomputed the profile for the compound in the MDAMB231 cell line with all profiles that had correlation coefficients above the threshold.

3. *Generating disease-specific drug response profiles:* In the last step (Figure 4C), we processed the outputs of the second step to extract a generalized response profile to a compound across the three TNBC cell lines. To do this, we performed independently steps 1 and 2 for QL-XII-47 in the other two TNBC cell lines available in the LINCS-L1000 project (BT20, HS578T) and averaged these profiles.

After summarizing the time points, concentrations, and cell lines for all the 205 compounds tested in TNBC cell lines, we ended with robust transcriptional response profiles of 152 (74.14%) compounds (Supplementary Table 1) that are specific for TNBC. The remaining 53 compounds were removed, as they did not show a stable response across time points, concentrations, and cell lines.

Having a single generalized disease-specific response profile for each compound allowed us to (1) generate compound combinations *in-silico* as candidates for the treatment of TNBC (Supplementary Table 2), (2) to rank compounds and compound combinations based on their predicted potential to reverse the disease state towards a healthy-like state (Supplementary Table 3 and 4), and to (3) interrogate the computed profiles using rank-based gene set enrichment analysis (GSEA) to predict possible mechanisms of action of the compounds when used to treat TNBC.

Our method to extract disease-specific drug response profiles is implemented in the '*retriever*' R package (see Data Availability section) and allows the computation of disease-specific transcriptional drug response signatures for different cancer types available in the LINCS-L1000 project.

Ranking compound candidates for the treatment of TNBC.

We used the 152 TNBC-specific transcriptional response profiles extracted with *retriever* to compute 11,476 response profiles to nonredundant drug combinations. To do so, we calculated

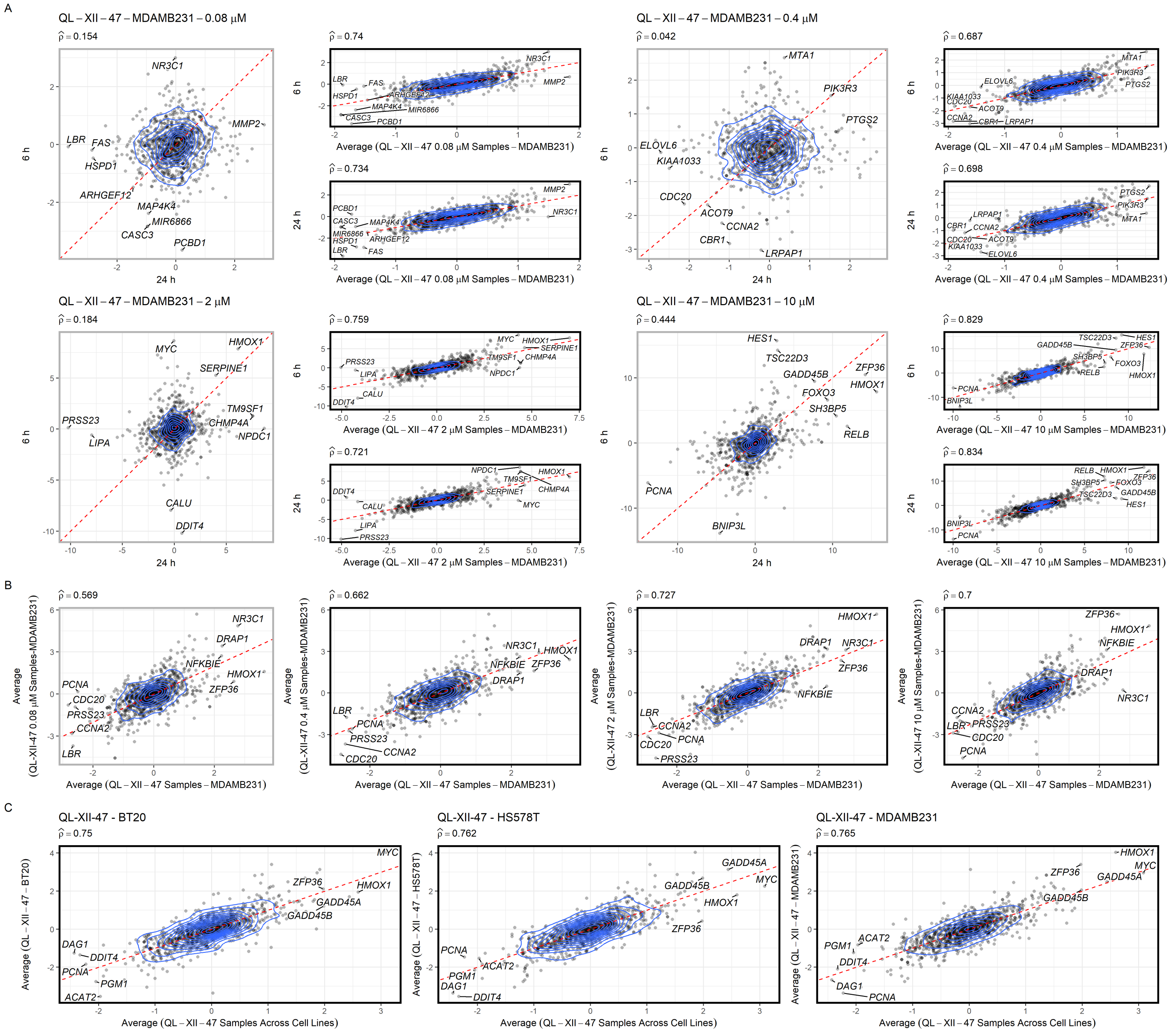

Fig. 4. Case example showing the construction of a single transcriptional response profile to a compound across three TNBC cell lines. Frames of the scatterplots displaying the relationship between the profiles and the averaged profiles are color coded, in black if the computed Spearman correlation coefficient ($\hat{\rho}$) is larger than 0.6 or in gray otherwise. **(A)** Generation of a time-consistent response profile. **(B)** Generation of a time and concentration-consistent response profile. **(C)** Generation of a time, concentration, and cell-line-consistent response profile.

the averaged effect sizes of Spearman correlation between the response profiles and the differential single-cell transcriptional signatures, and then ranked the compounds and compound combinations likely to specifically antagonize the TNBC-specific transcriptional signatures. Since Spearman correlation is a rank-based approach, this method can quantitatively identify inverse effects induced by drugs using the response profiles and disease-associated transcriptional signatures. Thus, we computed the Spearman coefficients, 95% confidence intervals, and p-values corrected for multiple testing (FDR) for all $11,628$ profiles (Supplementary Table 3 and 4). We then ranked the drug combinations by their correlation coefficients, from negative to positive. A negative value represents the expected inverse effect of the drug against the expression changes observed in the disease towards a healthy-like state.

We identified QL-XII-47 as the most promising compound to reverse the transcriptional profile of TNBC back to a healthy-like state, followed by Torin-2, Torin-1, QL-X-138, and WYE-125132. QL-XII-47 is a highly effective and selective Bruton tyrosine kinase (BTK) inhibitor that covalently modifies Cys481 of the protein. QL-XII-47 has an IC_{50} of 7 nM for inhibiting BTK kinase activity and induces a G1 cell cycle arrest in Ramos cells (B lymphocytes from Burkitt lymphoma), which is associated with significant degradation of the BTK protein. It was also shown that, at sub micromolar concentrations, QL-

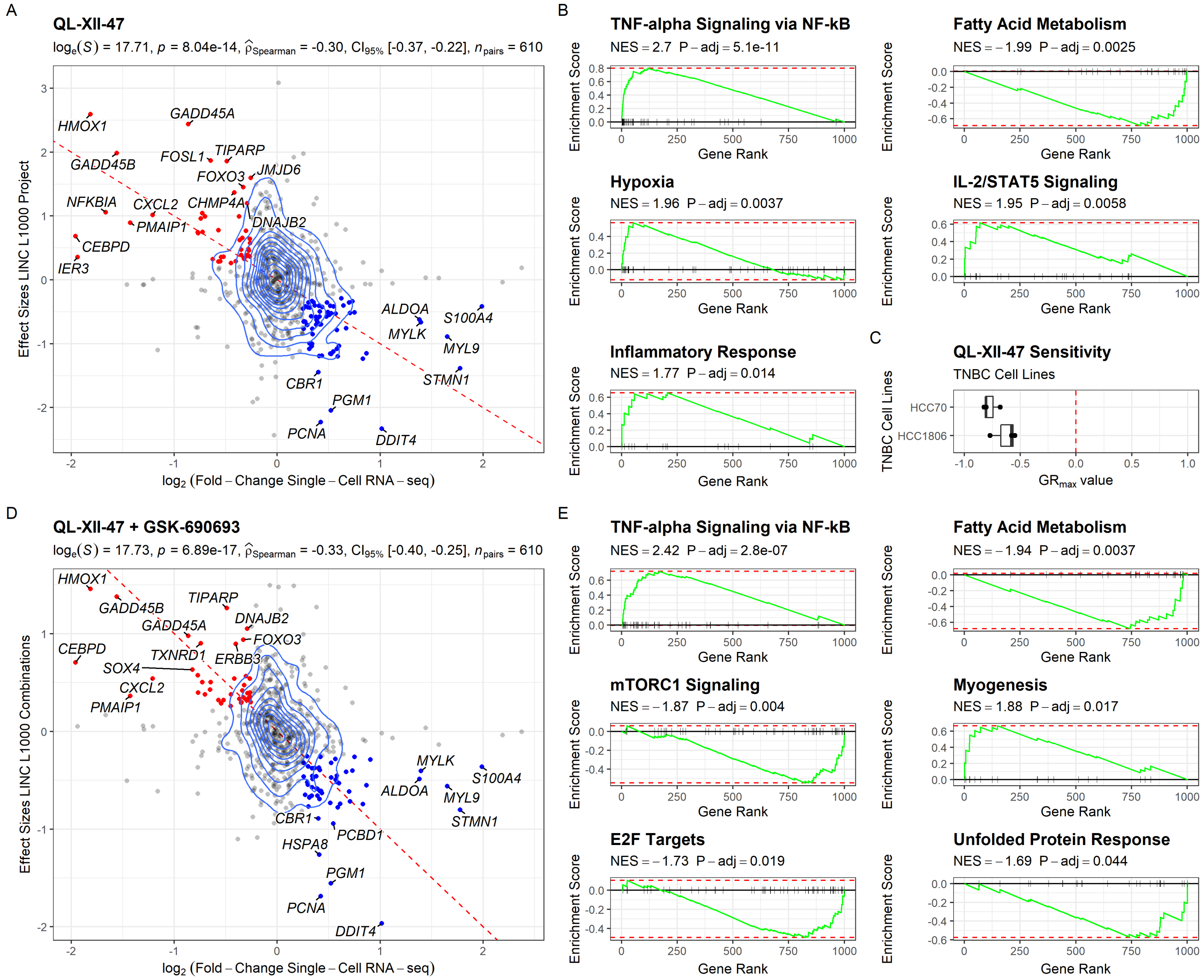

Fig. 5. Correlation analysis and mechanism of action prediction for the disease-specific drug response profiles in TNBC. **(A)** Spearman correlation analysis between the expression changes associated with TNBC and the disease-specific drug response signature of QL-XII-47. Each dot represents a gene. Dots are color coded, in red if they are expected to be upregulated by the drug, in blue if they are expected to be downregulated by the drug, and in gray if no significant change is expected. The red line represents the perfect match between both profiles. Density lines reflect the number of dots in each section of the plot. **(B)** Mechanisms of action predicted for QL-XII-47 in TNBC based on the enrichment of biological pathways in the disease-specific transcriptional drug response signature. **(C)** Independent sensitivity evaluation of the effect of QL-XII-47 in two other TNBC cell lines. **(D)** Spearman correlation analysis between the expression changes associated with TNBC and the combination signature of QL-XII-47 and GSK-690693. Each dot represents a gene. Dots are color coded as in (A). The red line represents the perfect match between both profiles. Density lines reflect the number of dots in each section of the plot. **(E)** Mechanisms of action predicted for the mixture of QL-XII-47 and GSK-690693 based on the enrichment of their combination drug response signature.

XII-47 inhibits the proliferation of B-cell lymphoma cell lines (50). Furthermore, other BTK inhibitors have shown to reduce TNBC cell viability (51). In addition, independent validation of sensitivity to QL-XII-47 has been tested before in two other TNBC cell lines (HC70 and HCC1806) that were not included in the LINCS-L1000 project. Low toxicity (lower than Torin-2) and more than 50% reduction of the growth rate ($\mathrm{GR}_{\max}$) was found for both cell lines (52).

Through enrichment analysis, we found that QL-XII-47 may act in TNBC through activation of the *TNF-alpha Signaling via NF-κB*, *Hypoxia-associated genes*, the *Inflammatory Response*, the *IL-2/STAT5 Signaling pathway*, and by deactivating genes involved in *Fatty Acid Metabolism* (FDR $<$ 0.05 in all cases, Fig 5A).

Ranking compound combination candidates for the treatment of TNBC. After testing the combinations of the disease-specific drug response profiles, we found that a combination of QL-XII-47 and GSK-690693—a pan-AKT kinase inhibitor that reduces tumor cell proliferation and induces tumor cell apoptosis (53, 54)—was the best performing drug combination to revert TNBC signatures, with an increase of 10% compared to monotherapy with QL-XII-47.

Gene set enrichment analysis predicted that this combination may

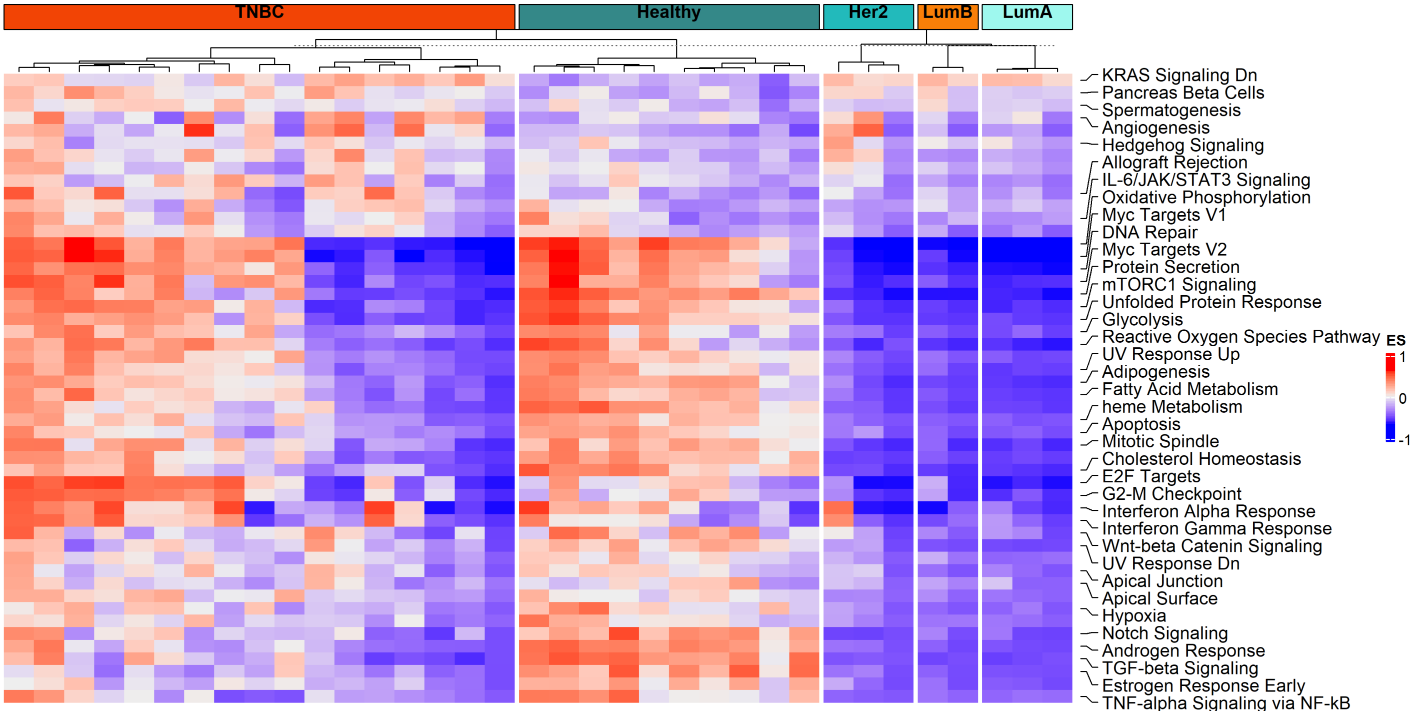

Fig. 6. Cancer hallmark signatures identified in triple-negative cells from individual breast cancer samples and triple-positive-like epithelial cells from control samples. Single-sample GSEA Enrichment Score (ES) of the pseudo-bulk profile computed for each donor included in the study. Labeled pathways are those that show a similar trend (either positive or negative ES) in the three data types. Dn: downregulation; Up: upregulation.

act through the activation of *TNF-alpha signaling via NF-κB* and *Myogenesis*, and the deactivation of *Fatty Acid Metabolism*, *mTORC1 Signaling*, *E2F targets*, and the *Unfolded Protein Response* (FDR < 0.05 in all cases). These pathways have previously been associated with the inhibition of triple-negative breast cancer growth and metastasis (55–57), and thus highlight the potential of our computational approach to prioritize drugs by integrating cell-type specific profiles obtained from single-cell RNA-seq data sets with disease-specific transcriptional drug response profiles.

Signatures impacted by QL-XII-47 and GSK-690693 in individual patients. Finally, we evaluated how combination treatment with QL-XII-47 and GSK-690693 might impact transcriptional profiles of triple-negative cancer cells at the individual patient level. We found that, for most patients, the signatures expected to be impacted by the drug combination were inversely correlated between healthy samples and cancer samples (Figure 6). For example, *TNF-alpha signaling via NF-κB* shows an inverse association in 84% (22 of 26) of patients, while *Hypoxia* shows an inverse association in 73% (19 of 26) patients. Note that, while we focused on triple-negative cells in this study, the breast cancer atlas we analyzed also included samples from receptor-positive cancers. However, we specifically applied *retriever* to triple-negative cells from these samples, aiming to identify drug combinations that would target the tumor's most difficult-to-treat cells. Interestingly, signatures in triple-negative cells from all of the included non-TNBC samples (8 of 8) showed strong inverse correlations to those from healthy samples. For the TNBC samples, 7 of 18 had strong inverse associations with the healthy sample cluster. This indicates that, when applied to single-cell data derived from a specific tumor type, the candidate drug combinations predicted by *retriever* are likely to benefit a relatively large number of patients.

Discussion

Single-cell RNA-seq provides an unprecedented resolution to characterize cellular heterogeneity in cancer. Compared with cell lines, the cells from tumors characterized by single-cell RNA-seq do not exhibit metabolic adaptation due to culturing. In advantage to bulk RNA-seq, single-cell RNA-seq can provide a more accurate signature of the changes observed in specific cells relevant to the disease under study(14, 58). Thus, leveraging the information provided by this new technology is important to accelerate the identification of new personalized treatments that target specific subpopulations of cells in a tissue (59).

However, just as important as characterizing the signatures of the disease in affected cells at the highest possible resolution is to characterize the response profiles of drugs that may help to reverse disease-associated transcriptional changes towards a healthy state. Large consortia including the LINCS-L1000 project provide transcriptional phenotypes at different time points after applying hundreds of compounds to multiple types of cell lines at different concentrations. Even though this information is valuable, its direct application to pathophysiological scenarios is limited due to the difficulty to extract robust response profile in these large data sets (60). Here, we introduced *retriever*, a tool to extract robust disease-specific drug response profiles from the LINCS-L1000 project. *retriever* mines the information associated to each cell line in the project to select subsets of cell lines that are surrogates of a defined disease, after which it extracts disease-specific response profiles. By doing this, *retriever* maximizes the robustness of the transcriptional signatures that are used to rank candidate compounds for the identification of new treatment strategies. With *retriever*, we hypothesize that if a signature is robust across multiple cell lines representing the same disease, it will likely also be more robust across individuals, as cell lines exhibit genetic and metabolic differences.

We showcased *retriever*'s strength by retrieving time-, dose-, and cell-type consistent transcriptional drug response signatures of TNBC cells from the LINCS-L1000 project and combining it with single-cell RNA transcriptome profiles obtained from a large single-cell breast cancer atlas, which we used to predict novel drug combinations against TNBC. This identified a combination of two kinase inhibitors predicted to be effective through together acting on important biological pathways in breast cancer, and thereby antagonize the TNBC-specific signature.

The prediction of drug combinations with *retriever* that we present here, although it is relatively simple and relies on the hypothesis of independent mechanisms of action for each compound, has been established as accurate (predicting the right directionality of the change) in analyses performed on profiles from the Connectivity Map project (28). Nevertheless, our computational approach only allows us to rank compounds and compounds combinations that are suitable for the development of treatments based on the negative correlation of the drug response profiles and the disease signatures. The topmost promising compounds and drug combinations still need to be validated experimentally to define the right concentration and evaluate if they indeed exhibit a synergistic mechanism of action. Their toxicity across cell lines and, after preclinical tests, potential adverse events in *in-vivo* models and cancer patients also need to be evaluated.

Thanks to the multiple cell lines available in the LINCS-L1000 project, our approach can be replicated in at least 13 other cancer types, including: *adult acute monocytic leukemia, adult acute myeloid leukemia, cecum adenocarcinoma, colon (adeno)carcinoma, endometrial adenocarcinoma, lung adenocarcinoma, large cell lung carcinoma, small cell lung carcinoma, melanoma, ovarian mucinous adenocarcinoma, prostate carcinoma*, and *triple-negative breast cancer* among other variants (such as local or metastatic tumors for some cancer types), as soon as single-cell RNA-seq data for those cancer types and healthy tissues become available. Considering the increase in single-cell RNA-seq studies being published since the development of the technique, we expect this to be possible for most cancer types represented in the LINCS-L1000 project in the near future (61).

Finally, our approach can also be applied to disease profiles derived from a single patient to recommend personalized treatment strategies. We envision this to be one of the potentially most impactful applications of *retriever* in computationally-informed precision medicine once single-cell transcriptional profiles become a standard in the clinic. In addition, while we have showcased *retriever*'s strength in predicting drug combinations for TNBC in general, the method can be applied to single-cell RNA-seq data derived from individual tumors, opening up the road for applications in precision medicine.

Data Availability

All the data and code required to replicate the analysis as well as the figures and tables are available at https://github.com/dosorio/L1000-TNBC. The *retriever* package is available at https://github.com/kuijjerlab/retriever. The following datasets were used to construct the single-cell RNA-seq breast atlas used in this study:

1. Tabula sapiens wild-type mammary gland. Data from (62) Quake, Stephen R., and Tabula Sapiens Consortium. *"The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors."* bioRxiv (2021). Accessed through: https://tabula-sapiens-portal.ds.czbiohub.org/

2. Wild-type data from seven individuals reported by (63) Nguyen, Quy H., *et al*. *"Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity."* Nature communications 9.1 (2018): 1-12. GEO accession code: GSE113197.

3. Wild-type data from five individuals reported by (64) Bhat-Nakshatri, Poornima, *et al*. *"A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells."* Cell Reports Medicine 2.3 (2021): 10021955. GEO accession code: GSE164898.

4. Triple-negative breast cancer data from 9 individuals from (34) Wu SZ, Al-Eryani G, Roden DL, Junankar S *et al*. *A single-cell and spatially resolved atlas of human breast cancers*. Nat Genet 2021 Sep;53(9):1334-134756. GEO accession code: GSE176078

5. Breast cancer data from 17 individuals from the Array-Express database accession code: E-MTAB-8107. Described in (65) Qian, Junbin, *et al*. *"A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling."* Cell research 30.9 (2020): 745-76257.

Acknowledgments

This work was supported by the Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (187615, M.L.K and D.O), the Norwegian Cancer Society (214871, M.L.K.), the Norwegian Research Council (313932, M.L.K), and the Marie Skłodowska-Curie Postdoctoral Scientia Fellows program from the Faculty of Medicine at the University of Oslo, co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant (801133, D.O).

Bibliography

1. L. Sleire, H. E. Forde, I. A. Netland, L. Leiss, B. S. Skeie, and P. O. Enger. Drug repurposing in cancer. *Pharmacol Res*, 124:74–91, 2017. ISSN 1096-1186 (Electronic) 1043-6618 (Linking). doi: 10.1016/j.phrs.2017.07.013.
2. R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead, and J. T. Dudley. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med*, 8(3):186–210, 2016. ISSN 1939-005X (Electronic) 1939-005X (Linking). doi: 10.1002/wsbm.1337.
3. T. P. Liu, Y. Y. Hsieh, C. J. Chou, and P. M. Yang. Systematic polypharmacology and drug repurposing via an integrated l1000-based connectivity map database mining. *R Soc Open Sci*, 5(11): 181321, 2018. ISSN 2054-5703 (Print) 2054-5703 (Linking). doi: 10.1098/rsos.181321.
4. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W. N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452 e17, 2017. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2017.10.049.
5. Y. Wang, J. Yella, and A. G. Jegga. Transcriptomic data mining and repurposing for computational drug discovery. *Methods Mol Biol*, 1903:73–95, 2019. ISSN 1940-6029 (Electronic) 1064-3745 (Linking). doi: 10.1007/978-1-4939-8955-3_5.
6. M. Ben Guebila, C. M. Lopes-Ramos, D. Weighill, A. R. Sonawane, R. Burkholz, B. Shamsaei, J. Platig, K. Glass, M. L. Kuijjer, and J. Quackenbush. Grand: a database of gene regulatory network models across human conditions. *Nucleic Acids Res*, 2021. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkab778.
7. E. S. Ozdemir, F. Halakou, R. Nussinov, A. Gursoy, and O. Keskin. Methods for discovering and targeting druggable protein-protein interfaces and their application to repurposing. *Methods Mol Biol*, 1903:1–21, 2019. ISSN 1940-6029 (Electronic) 1064-3745 (Linking). doi: 10.1007/978-1-4939-8955-3_1.
8. E. W. Su and T. M. Sanger. Systematic drug repositioning through mining adverse event data in clinicaltrials.gov. *PeerJ*, 5:e3154, 2017. ISSN 2167-8359 (Print) 2167-8359 (Linking). doi: 10.7717/peerj.3154.
9. M. Avalos-Moreno, A. Lopez-Tejada, J. L. Blaya-Canovas, F. E. Cara-Lupianez, A. Gonzalez-Gonzalez, J. A. Lorente, P. Sanchez-Rovira, and S. Granados-Principal. Drug repurposing for triple-negative breast cancer. *J Pers Med*, 10(4), 2020. ISSN 2075-4426 (Print) 2075-4426 (Linking). doi: 10.3390/jpm10040200.
10. A. Whitehead and D. L. Crawford. Variation in tissue-specific gene expression among natural populations. *Genome Biol*, 6(2):R13, 2005. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/gb-2005-6-2-r13.
11. Eyal Simonovsky, Ronen Schuster, and Esti Yeger-Lotem. Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety. *Bioinformatics*, 35(17):3028–3037, 2019.
12. H. M. Levitin, J. Yuan, and P. A. Sims. Single-cell transcriptomic analysis of tumor heterogeneity. *Trends Cancer*, 4(4):264–268, 2018. ISSN 2405-8025 (Electronic) 2405-8025 (Linking). doi: 10.1016/j.trecan.2018.02.003.
13. Consortium International Cancer Genome, T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolas, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P. A. Futreal, H. Aburatani, M. Bayes, D. D. Botwell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. J. Campbell, et al. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature08987.
14. S. S. Potter. Single-cell rna sequencing for the study of development, physiology and disease. *Nat Rev Nephrol*, 14(8):479–492, 2018. ISSN 1759-507X (Electronic) 1759-5061 (Linking). doi: 10.1038/s41581-018-0021-7.
15. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell rna sequencing. *Mol Cell*, 58(4):610–20, 2015. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2015.04.005.
16. D. Osorio, M. L. Kuijjer, and J. J. Cai. rpanglaodb: an r package to download and merge labeled single-cell rna-seq data from the panglaodb database. *Bioinformatics*, 2021. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btab549.
17. K. D. Zimmerman, M. A. Espeland, and C. D. Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun*, 12(1):738, 2021. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-021-21038-1.
18. J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. La Manno, M. A. Skinnider, and G. Courtine. Confronting false discoveries in single-cell differential expression. *Nat Commun*, 12(1):5692, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25960-2.
19. J. Caroli, M. Dori, and S. Bicciato. Computational methods for the integrative analysis of genomics and pharmacological data. *Front Oncol*, 10:185, 2020. ISSN 2234-943X (Print) 2234-943X (Linking). doi: 10.3389/fonc.2020.00185.
20. Y. Qiu, T. Lu, H. Lim, and L. Xie. A bayesian approach to accurate and robust signature detection on lincs l1000 data. *Bioinformatics*, 36(9):2787–2795, 2020. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btaa064.
21. T. J. Gonda and R. G. Ramsay. Directly targeting transcriptional dysregulation in cancer. *Nat Rev Cancer*, 15(11):686–94, 2015. ISSN 1474-1768 (Electronic) 1474-175X (Linking). doi: 10.1038/nrc4018.
22. N. Chatterjee and T. G. Bivona. Polytherapy and targeted cancer drug resistance. *Trends Cancer*, 5(3):170–182, 2019. ISSN 2405-8025 (Electronic) 2405-8025 (Linking). doi: 10.1016/j.trecan.2019.02.003.
23. M. Jeon, S. Kim, S. Park, H. Lee, and J. Kang. In silico drug combination discovery for personalized cancer therapy. *BMC Syst Biol*, 12(Suppl 2):16, 2018. ISSN 1752-0509 (Electronic) 1752-0509 (Linking). doi: 10.1186/s12918-018-0546-1.
24. H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng. Drugcombdb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res*, 48(D1):D871–D881, 2020. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkz1007.
25. R. Celebi, O. th Bear Don't Walk, R. Movva, S. Alpsoy, and M. Dumontier. In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data. *Sci Rep*, 9(1):8949, 2019. ISSN 2045-2322 (Electronic) 2045-2322 (Linking). doi: 10.1038/s41598-019-45236-6.
26. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795): 1929–35, 2006. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1132939.
27. Alexander V. Pickering. ccmap: Combination connectivity mapping. *Bioconductor*, 2021. doi: 10.18129/B9.bioc.ccmap.
28. Alexander V. Pickering. ccdata: Data for combination connectivity mapping (ccmap) package. *Bioconductor*, 2021. doi: 10.18129/B9.bioc.ccdata.
29. K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J. C. Marugan, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S. E. Hunt, I. Isley GR, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2021. *Nucleic Acids Res*, 49(D1):D884–D891, 2021. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkaa942.
30. Y. Hao, S. Hao, E. Andersen-Nissen, 3rd Mauck, W. M., S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587 e29, 2021. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2021.04.048.
31. D. Osorio and J. J. Cai. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell rna-sequencing data quality control. *Bioinformatics*, 37(7):963–967, 2021. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btaa751.
32. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods*, 16(12):1289–1296, 2019. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/s41592-019-0619-0.
33. E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nat Biotechnol*, 2018. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/nbt.4314.
34. S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, T. Wang, L. Larsson, D. Kaczorowski, N. I. Weisenfeld, C. R. Uytingco, J. G. Chew, Z. W. Bent, C. L. Chan, V. Gnanasambandapillai, C. A. Dutertre, L. Gluch, M. N. Hui, J. Beith, A. Parker, E. Robbins, D. Segara, C. Cooper, C. Mak, B. Chan, S. Warrier, F. Ginhoux, E. Millar, J. E. Powell, S. R. Williams, X. S. Liu, S. O'Toole, E. Lim, J. Lundeberg, C. M. Perou, and A. Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet*, 53(9):1334–1347, 2021. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/s41588-021-00911-1.
35. Sunny Z Wu, Daniel L Roden, Chenfei Wang, Holly Holliday, Kate Harvey, Aurélie S Cazet, Kendelle J Murphy, Brooke Pereira, Ghamdan Al-Eryani, Nenad Bartonicek, et al. Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *The EMBO journal*, 39(19):e104063, 2020.
36. J. Alquicira-Hernandez and J. E. Powell. Nebulosa recovers single cell gene expression signals by kernel density estimation. *Bioinformatics*, 2021. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btab003.
37. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biol*, 16:278, 2015. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-015-0844-5.
38. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-014-0550-8.
39. A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Res*, 44(8):e71, 2016. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkv1507.
40. C. Spearman. The proof and measurement of association between two things. by c. spearman, 1904. *Am J Psychol*, 100(3-4):441–71, 1987. ISSN 0002-9556 (Print) 0002-9556 (Linking).
41. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1):289–300, 1995.
42. Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo

Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6): 417–425, 2015.

43. Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel JB Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, et al. Gene set knowledge discovery with enrichr. *Current protocols*, 1(3):e90, 2021.
44. K. W. Evans, E. Yuca, S. S. Scott, M. Zhao, N. Paez Arango, C. X. Cruz Pico, T. Saridogan, M. Shariati, C. A. Class, C. A. Bristow, C. P. Vellano, X. Zheng, A. M. Gonzalez-Angulo, X. Su, C. Tapia, K. Chen, A. Akcakanat, B. Lim, D. Tripathy, T. A. Yap, M. E. Di Francesco, G. F. Draetta, P. Jones, T. P. Heffernan, J. R. Marszalek, and F. Meric-Bernstam. Oxidative phosphorylation is a metabolic vulnerability in chemotherapy-resistant triple negative breast cancer. *Cancer Res*, 2021. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). doi: 10.1158/0008-5472.CAN-20-3242.
45. D. Horiuchi, L. Kusdra, N. E. Huskey, S. Chandriani, M. E. Lenburg, A. M. Gonzalez-Angulo, K. J. Creasman, A. V. Bazarov, J. W. Smyth, S. E. Davis, P. Yaswen, G. B. Mills, L. J. Esserman, and A. Goga. Myc pathway activation in triple-negative breast cancer is synthetic lethal with cdk inhibition. *J Exp Med*, 209(4):679–96, 2012. ISSN 1540-9538 (Electronic) 0022-1007 (Linking). doi: 10.1084/jem.20111512.
46. O. K. Provance, E. S. Geanes, A. J. Lui, A. Roy, S. M. Holloran, S. Gunewardena, C. R. Hagan, S. Weir, and J. Lewis-Wambi. Disrupting interferon-alpha and nf-kappab crosstalk suppresses ifitm1 expression attenuating triple-negative breast cancer progression. *Cancer Lett*, 514:12–29, 2021. ISSN 1872-7980 (Electronic) 0304-3835 (Linking). doi: 10.1016/j.canlet.2021.05.006.
47. V. S. Jamdade, N. Sethi, N. A. Mundhe, P. Kumar, M. Lahkar, and N. Sinha. Therapeutic targets of triple-negative breast cancer: a review. *Br J Pharmacol*, 172(17):4228–37, 2015. ISSN 1476-5381 (Electronic) 0007-1188 (Linking). doi: 10.1111/bph.13211.
48. L. Gerratana, D. Basile, G. Buono, S. De Placido, M. Giuliano, S. Minichillo, A. Coinu, F. Martorana, I. De Santo, L. Del Mastro, M. De Laurentiis, F. Puglisi, and G. Arpino. Androgen receptor in triple negative breast cancer: A potential target for the targetless subtype. *Cancer Treat Rev*, 68:102–110, 2018. ISSN 1532-1967 (Electronic) 0305-7372 (Linking). doi: 10.1016/j.ctrv.2018.06.005.
49. M. Oshi, S. Newman, Y. Tokumaru, L. Yan, R. Matsuyama, I. Endo, and K. Takabe. Inflammation is associated with worse outcome in the whole cohort but with better outcome in triple-negative subtype of breast cancer patients. *J Immunol Res*, 2020:5618786, 2020. ISSN 2314-7156 (Electronic) 2314-7156 (Linking). doi: 10.1155/2020/5618786.
50. H. Wu, W. Wang, F. Liu, E. L. Weisberg, B. Tian, Y. Chen, B. Li, A. Wang, B. Wang, Z. Zhao, D. W. McMillin, C. Hu, H. Li, J. Wang, Y. Liang, S. J. Buhrlage, J. Liang, J. Liu, G. Yang, J. R. Brown, S. P. Treon, C. S. Mitsiades, J. D. Griffin, Q. Liu, and N. S. Gray. Discovery of a potent, covalent btk inhibitor for b-cell lymphoma. *ACS Chem Biol*, 9(5):1086–91, 2014. ISSN 1554-8937 (Electronic) 1554-8929 (Linking). doi: 10.1021/cb4008524.
51. R. Campbell, G. Chong, and E. A. Hawkes. Novel indications for bruton's tyrosine kinase inhibitors, beyond hematological malignancies. *J Clin Med*, 7(4), 2018. ISSN 2077-0383 (Print) 2077-0383 (Linking). doi: 10.3390/jcm7040062.
52. Sameer S Chopra, Anne Jenney, Adam Palmer, Mario Niepel, Mirra Chung, Caitlin Mills, Sindhu Carmen Sivakumaren, Qingsong Liu, Jia-Yun Chen, Clarence Yapp, et al. Torin2 exploits replication and checkpoint vulnerabilities to cause death of pi3k-activated triple-negative breast cancer cells. *Cell systems*, 10(1):66–81, 2020.
53. D. A. Altomare, L. Zhang, J. Deng, A. Di Cristofano, A. J. Klein-Szanto, R. Kumar, and J. R. Testa. Gsk690693 delays tumor onset and progression in genetically defined mouse models expressing activated akt. *Clin Cancer Res*, 16(2):486–96, 2010. ISSN 1557-3265 (Electronic) 1078-0432 (Linking). doi: 10.1158/1078-0432.CCR-09-1026.
54. D. S. Levy, J. A. Kahana, and R. Kumar. Akt inhibitor, gsk690693, induces growth inhibition and apoptosis in acute lymphoblastic leukemia cell lines. *Blood*, 113(8):1723–9, 2009. ISSN 1528-0020 (Electronic) 0006-4971 (Linking). doi: 10.1182/blood-2008-02-137737.
55. W. Wang, R. Zhang, X. Wang, N. Wang, J. Zhao, Z. Wei, F. Xiang, and C. Wang. Suppression of kif3a inhibits triple negative breast cancer growth and metastasis by repressing rb-e2f signaling and epithelial-mesenchymal transition. *Cancer Sci*, 111(4):1422–1434, 2020. ISSN 1349-7006 (Electronic) 1347-9032 (Linking). doi: 10.1111/cas.14324.
56. A. Sulaiman, S. McGarry, K. M. Lam, S. El-Sahli, J. Chambers, S. Kaczmarek, L. Li, C. Addison, J. Dimitroulakos, A. Arnaout, C. Nessim, Z. Yao, G. Ji, H. Song, S. Liu, Y. Xie, S. Gadde, X. Li, and L. Wang. Co-inhibition of mtorc1, hdac and esr1alpha retards the growth of triple-negative breast cancer and suppresses cancer stem cells. *Cell Death Dis*, 9(8):815, 2018. ISSN 2041-4889 (Electronic). doi: 10.1038/s41419-018-0811-7.
57. R. Camarda, A. Y. Zhou, R. A. Kohnz, S. Balakrishnan, C. Mahieu, B. Anderton, H. Eyob, S. Kajimura, A. Tward, G. Krings, D. K. Nomura, and A. Goga. Inhibition of fatty acid oxidation as a therapy for myc-overexpressing triple-negative breast cancer. *Nat Med*, 22(4):427–32, 2016. ISSN 1546-170X (Electronic) 1078-8956 (Linking). doi: 10.1038/nm.4055.
58. Shoval Lagziel, Eyal Gottlieb, and Tomer Shlomi. Mind your media. *Nature Metabolism*, 2(12): 1369–1372, 2020.
59. Julia E Wiedmeier, Pawan Noel, Wei Lin, Daniel D Von Hoff, and Haiyong Han. Single-cell sequencing in precision medicine. In *Precision Medicine in Cancer Therapy*, pages 237–252. Springer, 2019.
60. Zhaleh Safikhani, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, Nicolai J Birkbak, Christos Hatzis, Leming Shi, et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Research*, 5, 2016.
61. Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. *Database*, 2020, 2020.
62. The Tabula Sapiens Consortium and Stephen R Quake. The tabula sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. *bioRxiv*, 2021.
63. Q. H. Nguyen, N. Pervolarakis, K. Blake, D. Ma, R. T. Davis, N. James, A. T. Phung, E. Willey, R. Kumar, E. Jabart, I. Driver, J. Rock, A. Goga, S. A. Khan, D. A. Lawson, Z. Werb, and K. Kessenbrock. Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity. *Nat Commun*, 9(1):2028, 2018. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-018-04334-1.
64. P. Bhat-Nakshatri, H. Gao, L. Sheng, P. C. McGuire, X. Xuei, J. Wan, Y. Liu, S. K. Althouse, A. Colter, G. Sandusky, A. M. Storniolo, and H. Nakshatri. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep Med*, 2(3):100219, 2021. ISSN 2666-3791 (Electronic) 2666-3791 (Linking). doi: 10.1016/j.xcrm.2021.100219.
65. J. Qian, S. Olbrecht, B. Boeckx, H. Vos, D. Laoui, E. Etlioglu, E. Wauters, V. Pomella, S. Verbandt, P. Busschaert, A. Bassez, A. Franken, M. V. Bempt, J. Xiong, B. Weynand, Y. van Herck, A. Antoranz, F. M. Bosisio, B. Thienpont, G. Floris, I. Vergote, A. Smeets, S. Tejpar, and D. Lambrechts. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res*, 30(9):745–762, 2020. ISSN 1748-7838 (Electronic) 1001-0602 (Linking). doi: 10.1038/s41422-020-0355-0.

Supplementary Material:

Supplementary Figures:

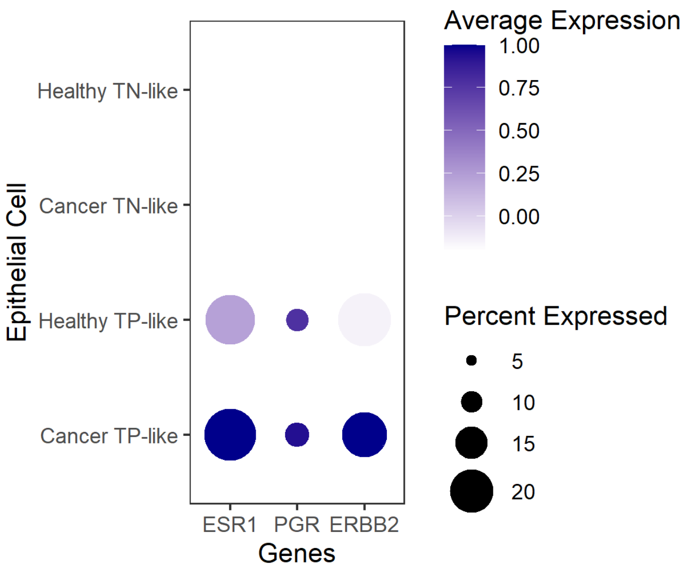

Fig. S1. Comparisons of *ESR1*, *PGR*, and *ERBB2* expression levels across epithelial cells in cancer and healthy tissues. TP stands for Triple-Positive, and TN for Triple-Negative.

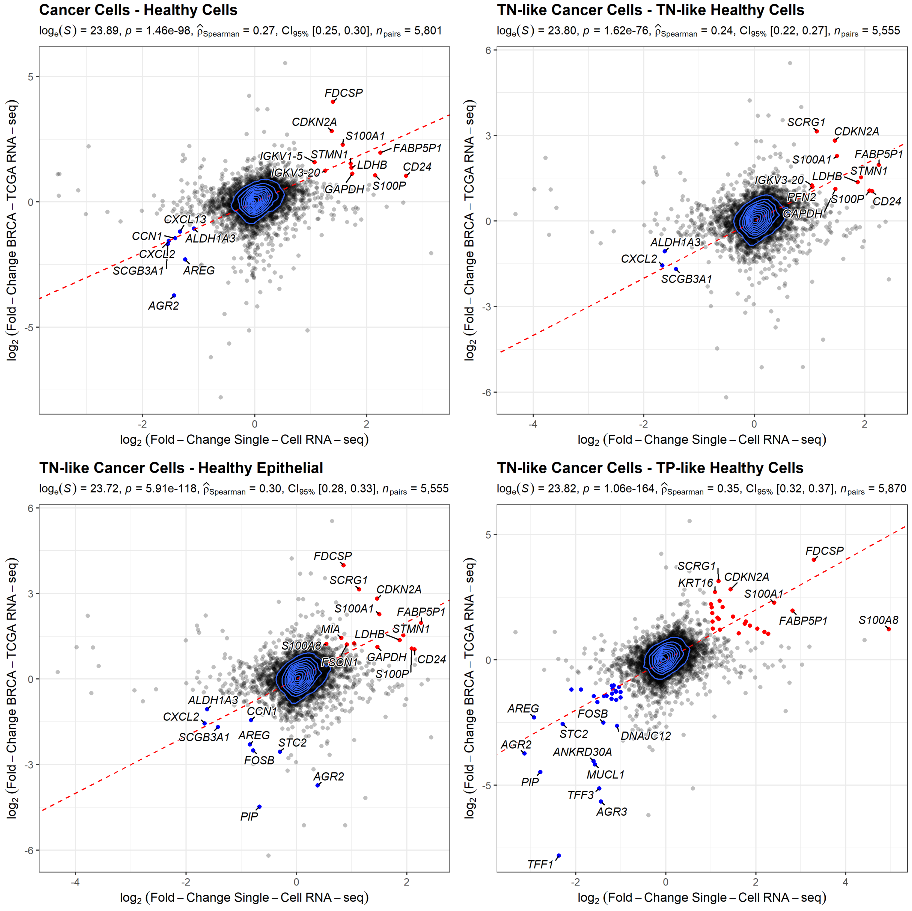

Fig. S2. Comparisons of the transcriptional changes associated with TNBC at the single-cell and tissue level across different subpopulations of cells. Each dot represents a gene. Dots are color coded, in red if the $\log_2$ fold-change is larger than 1 and in blue if the $\log_2$ fold-change is smaller than -1. TP = Triple-Positive, TN = Triple-Negative.

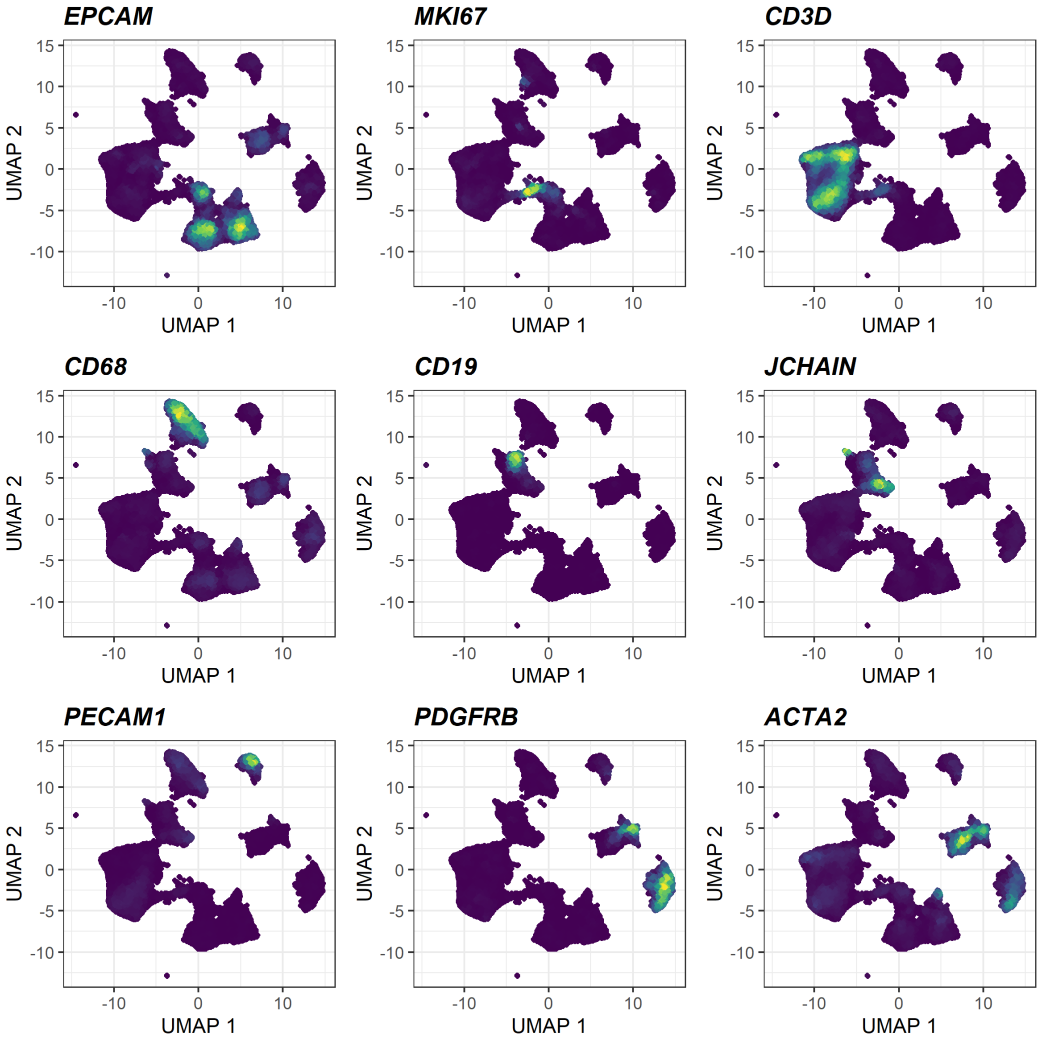

Fig. S3. Expression of markers for epithelial cells (*EPCAM*), proliferating cells (*MKI67*), T cells (*CD3D*), myeloid cells (*CD68*), B cells (*MS4A1*), plasmablasts (*JCHAIN*), endothelial cells (*PECAM1*), mesenchymal cells (fibroblasts/perivascular-like cells; *PDGFRB*), and muscular cells (*ACTA2*).

Supplementary Tables:

Table S2: (CSV File) Computed robust disease-specific response profiles to 152 compounds across three triple-negative breast cancer cell lines (TNBC)

Table S3: (CSV File) Computed $11,476$ compound combinations response profiles across three triple-negative breast cancer cell lines (TNBC)

Table S4: (CSV File) Ranked compounds based on their predicted potential to reverse the disease state (TNBC) towards a healthy state

Table S5: (CSV File) Ranked compound combinations based on their predicted potential to reverse the disease state (TNBC) towards a healthy state

Cancer Hallmark	**FDR**	**Leading Genes**
Oxidative Phosphorylation	1.76×10^{-4}	*LDHB, IDH2, SLC25A5, ATP5F1C, COX6C, PHB2, ATP6V0C,* and UQCRH
Interferon Alpha Response	6.10×10^{-3}	*CD74, IFITM1, IFI27,* and *LY6E*
Myc Targets	4.24×10^{-2}	*RPLP0, PABPC1, PHB2,* and *PPIA*
TNFα Signaling via NF-κB	2.95×10^{-19}	*EDN1, BTG1, GADD45B, TSC22D1, CEBPD, CXCL1, SOD2, AREG, CXCL2, NFKBIA, MARCKS, KLF6, MAFF, CCL2, FOSB, CCN1, IL6ST, IER5, ATF3,* and *IER3*
Estrogen Response	6.54×10^{-8}	*UGCG, KRT18, ELOVL5, STC2, HSPB8, SLC39A6, TFF3, PMAIP1, TFF1, IL6ST,* and *AREG*
UV Response Up	6.54×10^{-8}	*NFKBIA, BTG1, HNRNPU, HMOX1, FOSB, TMBIM6, IL6ST, SOD2, CXCL2,* and *ATF3*
Apoptosis	6.54×10^{-8}	*WEE1, KRT18, GADD45B, TNFRSF12A, PMAIP1, HMOX1, EMP1, SOD2, ATF3,* and *IER3*
Hypoxia	3.45×10^{-7}	*KLF6, BTG1, HSPA5, STC2, MAFF, HMOX1, ADM, CCN1, ATF3,* and *IER3*
Unfolded Protein Response	9.36×10^{-5}	*HSPA5, STC2, DNAJB9, CCL2, ATF3*, and *ATF4*
IL-6/JAK/STAT3 Signaling	2.59×10^{-4}	*TNFRSF12A, HMOX1, CXCL1, CXCL13,* and *IL6ST*
Inflammatory Response	1.57×10^{-3}	*NFKBIA, EDN1, KLF6, CXCL8, CCL2*, and *ADM*
Androgen Response	4.62×10^{-3}	*TSC22D1, ELOVL5, DNAJB9,* and *SLC38A2*

Table S1. Hallmark MSigDB signatures associated with the differentially expressed genes observed in triple-negative breast cancer (TNBC). False Discovery Rate (FDR) computed using Gene Set Enrichment Analysis (GSEA).